COMPUTATIONAL MINING OF GENOMIC AND PROTEOMIC DATA TO GAIN INSIGHT FOR ALZHEIMER'S DISEASE (AD)

SYNOPSIS

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

Bioinformatics

By

ASHWANI KUMAR

Enrolment No. 136506

Under the supervision of

Dr. Tiratha Raj Singh

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT-173234, HP, INDIA

August, 2017

COMPUTATIONAL MINING OF GENOMIC AND PROTEOMIC DATA TO GAIN INSIGHT FOR ALZHEIMER'S DISEASE (AD)

ASHWANI KUMAR

INTRODUCTION

In a population that is growing older and older, it is difficult to underestimate the effect of age-related neurodegenerative disorders such as Alzheimer's disease (AD) [1]. AD is the major cause of death among people of America along with HIV; breast cancer; prostate cancer; stroke and heart disease; the percentage of death by different type of syndrome is shown in Figure 1. In particular, the absence of treatment options to effectively treat, or may be slow down the progression of these disorders implies that individuals diagnosed with AD face a long and devastating disease with an unavoidably negative outcome. This introduction will first focus on the clinical and neuropathological characteristics of AD. Subsequently, advances within the understanding of alleged molecular mechanisms principal neurodegenerative events, primarily based on studies of sporadic and rare familial variants of AD, will be discussed. As mentioned, effective treatment regimes remain absent, despite enormous research efforts during the last four decades [2]. Therefore, the final part of this introduction will focus on new research strategies for biomarkers identification that might help to provide insight in the biological mechanisms that contribute to AD-associated neurodegeneration.

AD remains poorly understood and represents a significant and growing scientific, social and economic challenge to many countries. It may be the ultimate challenge because there is no vaccine, no known prevention strategy, and no cure [3], and is thus a perfect focus for this work.

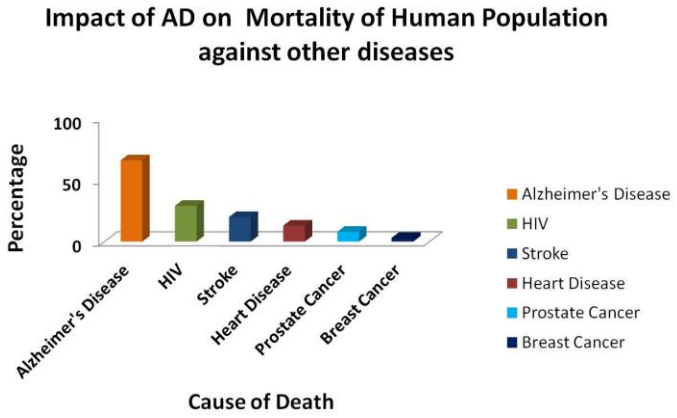

Figure 1: Percentage changes in selected causes of death (all ages) between 2010-2016 among people of America

AD is an irreversible, progressive brain disease that, along with related dementias, affects 44 millions of people worldwide [4] and in India it is about 0.4 million [5] from world health organization (WHO) report of 2016 . It gradually destroys brain function, resulting in cognitive feature decline (e.g. memory loss, language difficulty), worsening of some motor functions, behavioral and psychiatrically disorder (e.g. depression, delusion, and agitation), and decline in lively status (e.g. ability to engage in activities of daily living and self-care) [6]. Eventually, a person with the disease may no longer recognize family and friends and become completely dependent on others for help with even for the basic activities of daily living, such as eating and bathing.

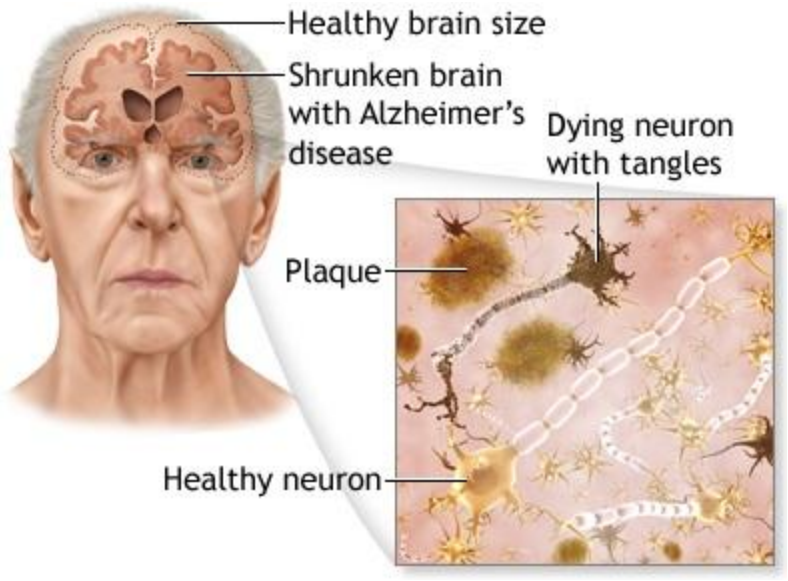

Figure 2: Alzheimer's disease and its major hallmarks (Source: Stem cell Tech)

AD is identified by the presence of hallmarks like extracellular senile plaques composed of amyloid-β (Aβ) peptide and intraneuronal neurofibrillary tangles (NFT) containing hyperphosphorylated tau protein [7] (Figure 2). The amyloid and the tau hypotheses consider these proteins as inducers and key players of the disease. Genes have a diversed influence on developing AD, starting from the autosomal dominant inheritance within the familial forms (1-5% of cases) to the polygenic background in late-onset (>65 years of age) sporadic AD (≥95% of cases). In addition to the genetic component, the risk for developing AD, the age at onset and also the course of the disease are influenced by many alternative factors including living style, atmosphere and therapeutic conditions [8]. The hereditary part however, might be of major importance, since according to clone studies, a major part of the risk for sporadic AD is genetically determined.

The amyloid precursor protein (APP), presenilin 1 (PSEN1) and presenilin 2 (PSEN2) genes are presently known to be concerned in the familial types of AD [9]. Identification and characterization of dominant mutations of those genes was useful for the understanding of the biological mechanisms resulting in increased Aβ accumulation and senile plaques formation. In contrast with the familial AD, the inflicting factors of the Aβ accumulations and other

pathological mechanisms stay largely unclear in the sporadic form. Advance analysis for sporadic AD suggests that different conditional sets of genes may converge on the pathological processes that underlie the illness. However, to this point solely the apolipoprotein E (APOE) gene has been definitively related with the risk for AD [10]. APOE is concerned in macromolecule transport and metabolism. Moreover, it plays a particular role within the central nervous system, as well as neural development, regeneration and other neurodegenerative processes. The polymorphism of the APOE gene determines three isoforms of APOE protein (ε2, ε3, ε4). APOE ε4 isoform increase the possibility of Aβ production [11].

The amyloid cascade hypothesis has been the principal hypothesis of molecular mechanisms underlying the pathological process of AD. According to this model, the fundamental cause of AD is the accumulation and aggregation of the Aβ peptide in the brain. The amyloid cascade hypothesis is widely accepted, but does not explain all aspects of AD pathology [12], and the major role of other infective mechanisms is also additionally supposable as well as inflammation, oxidative stress, lipid dysfunction, communication deficits and neuronal pathology .

Since a large part of the herediatary factors continues to be poorly understood and remains unanswered. The aim of our work was to contribute to this field investigating the biomarker(s) of genes and proteins related to AD. A biomarker of any disease would ideally allow for the measurement of the pathological process associated with a disease, whether or not it was linked to any direct pathology. Therefore a biomarker could be used to determine whether the disease is or is not present and be part of the clinical diagnosis process, and whether any potential therapies for the treatment of the disease are effective. In AD a biomarker could be based in the cerebrospinal fluid, or a blood-based gene, protein or image etc. The biological marker(s) could potentially lead to the development of new drug treatments and an early and accurate diagnosis of AD [13].

Keeping in view the gaps existing in AD research, the current study was undertaken with the following objectives:

Objective 1: Development of Comprehensive Biomarkers Database for AD: ABCD.

Objective 2: Classification and system level enrichment analysis of candidate geneset of AD.

Objective 3: Molecular characterization of specific chemical group of compounds to find inhibitors for specific target protein Acetylcholinesterase (AChE).

MATERIALS AND METHODS

Data Collection: AD genes data were collected from various authenticated freely available online resources like ensemble gene, Alzgene , Gencard, Genopedia and NCBI. Protein data (sequences and structures) obtained from Uniprot, human protein reference database (HPRD), Protein information resources (PIR) and PDB. Various molecular data like SNPs, miRNA, transcription factors (TF) etc. from dbSNP, miRbase and NCBI respectively. Clinical trials of AD were obtained from ClinicalTrial.gov, drugs candidates for AD inhibition were taken from Pubchem and Chembank. Reference pathway data were received from KEGG and Reactome. MRI scanned data of AD affected brain image data were collected from Alzheimer's Disease Neuroimaging (ADNI) on request [14]. Literature about AD pathogenesis got from Pubmed, Google Scholar, OMIM, MESH and National Institute on Ageing (NIA).

Tools**: (i)** HTML and PHP are two scripting language use for construction of frontend or graphical user interface (GUI) and connectivity purpose. Mysql is a relational language used for construction of database and Apache server was used for maintaining strong connection for data storage and and data retrieval.

(ii) Classification of geneset data through machine learning techniques was carried out through RapidMiner studio 6.2.0 [15], Weka 3.6.9 [16] and enrichment analysis for functional gene classification through David tool [17]. Classified genes from DT used as query with reference associated with AD obtained from Genopedia and Animal study. Associated partners of genes which code for proteins took from STRING database, STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations [18]. Later, construct network and visualize important topological parameters of AD pathways and its association with other diseases on basis of enrichment score in Cytoscape of JEPETTO plugin [19].

(iii) Lastly**,** For the purpose of characterization of Acetylcholnesterase (AChE) inhibitors, molecular docking was performed using AutoDock version 4.2.6 [20]. The active site of AChE was identified using the Catalytic Site Atlas database [21]. Subsequently, Molecular Dynamics

(MD) simulation was carried out using Desmond v3.7 and the results were rendered using Schrodinger's maestro v9.7 vizualizer interface [22] and further in-vitro two test were performed first by acetlcholine assay to check IC_{50} of six compounds and also cytotoxicity assay of same six compounds were evaluated against $Hgcl_2$ to test their potential to inhibit the process of toxicity in neuronic cell.

Methodology for objective 1

ABCD is a relational database implemented in MySQL language with a web interface that was developed using phpMyAdmin platform and hosted on Apache http server, which allows user to access and retrieve information in easy way without putting much effort and store various integrated molecular data related to AD in it.

Methodology for objective 2

Building a decision tree is a popular machine learning approach to classify the gene data. Decision tree generates rules for classification. A tree is represented by the set of nodes, leaves or branches. The root node is the attribute from which classification process starts, and the internal node corresponds to each of the questions about the particular attribute of the problem. The branches coming out of the each node are labeled with possible values of attribute. Each leaf node corresponds to a decision [23]. Decision tree model build successively with the set of AD genes from different sources, which is followed by preprocessing of the dataset including converting data into numeric values. Using geneset of 2111 gene, we use RapidMiner statistical tool and Weka tool for datamining task. Further, we validated our model by using 10-fold cross-validation application operator. Gene data includes the descriptions like Gene Name, Association Score, Chromosomal position, Minimental state examination (MMSE) score etc [24]. Enrichment analysis was also performed for this geneset to group genes on the basis of similarity using clustering algorithm. For more tightly associated gene in each group stringency was maintained by keeping the kappa threshold 0.3 as anything below this threshold have great chance to be a noise. Finally; GO annotation was described through biological process and molecular functions for the same.

The next goal of our study was to understand the regulatory mechanism of AD by enrichment study of genes associated to disease. We have gathered information concerning 1669

human genes referenced from PubMed was connected with AD. JEPETTO plugin in *Cytoscape* provides two forms of analysis. Firstly, the enrichment associate analysis finds pathways that measure powerfully related to a query geneset in term of an interaction network supported XD-score [25]. XD-score uses stochastic algorithmic to find a distance between the mapped genes and also the pathways within the molecular interaction network. Secondly, topology analysis is performed to seek out an analogous set of topological options shared by the pathways. Each analysis was performed on an individual basis [26]. Practical interpretation of geneset is feasible by activity associate enrichment analysis and by that specializes in pathways with highest scores. One might additionally generate the network of interactions for every pathway of interest and analyze the overlap with the geneset whereas the protein encoding genes may be additional as a path enlargement. KEGG, Reactome and GO reference databases were used for comparison purpose. The topological analysis was performed to check the interaction pattern between the AD pathways and input geneset, it was anticipated that the proposed enrichment study could help in the disease regulatory processes.

Methodology for objective 3

Molecular Docking

Molecular docking was performed using AutoDock version 4.2 [27]. AutoDock 4.2 retains the functionality of earlier versions, but adds the options of using a genetic algorithm (GA) for global searching, a local search (LS) method to perform energy minimization, or a combination of both. The structure of AChE for molecular docking studies was retrieved from Protein Data Bank (pdb id: 1B41) [28]. The protein preparation steps involved removing water molecules, metal ions and cofactors. Partial Kollman charges were applied to the protein after adding the hydrogen ions. The molecules used for molecular docking were Begacestat, Quercetin, Ascorbic acid, Donepezil, Caffeine and Gallic acid were appropriately prepared with molecule properties were given in Table1. The active site of AChE was identified using the Catalytic Site Atlas database. The list of residues that are listed as active site or catalytic site residues were Gly122, Gly121, Tyr133, Glu202, Ser 203, Ala204, His447, Glu450, and Glu334. A gridbox of dimension 90x90x90 Å was constructed with the α-carbon of Ser203 as its center. Autogrid was performed with the above settings to generate appropriate receptor and molecule maps. AutoDock parameter set and distance-dependent dielectric functions were used in the calculation

of the van der Waals and the electrostatic terms, respectively. Docking simulations were performed using genetic algorithm, where the initial position, orientation, and torsions of the molecule molecules were set randomly. The other parameters used for genetic algorithm were number of individual in a population (50), maximum number of energy evaluation (25000000), rate of gene mutation (0.2) and number of GA run (50). Docking calculations were carried out based on the binding free energy and their inhibition constant (K_i) for the following molecules Donepezil, Begacestat, Ascorbic acid, Caffiene ,Gallic acid and Quercetin towards the AChE. Finally, the best confirmation of the docked molecules was picked from the cluster analysis of docked conformations.

Molecular Dynamics simulation

Molecular Dynamics (MD) simulation was carried out using Desmond v3.7 and the results were rendered using Schrodinger's maestro v9.7 vizualizer interface. Desmond is a high performance MD simulations program available for biomolecular systems using novel parallel algorithms as scalability and numerical techniques to achieve accuracy [22]. Calculated scores, helps in analyzing the best binding pose. Each generated conformer was subjected to a full minimization in the gas phase with the Optimized Potential for Liquid Simulations (OPLS2005) all-atom force field [29]. Individual particle motions and its interaction with other particles were measured through simulation studies. Initially, the co-ordinates for MD simulations were taken from the protein-molecule complex. The Simple Point Charge (SPC) solvent model was used with each water box measuring orthorhombic dimensions of 10x10x10Å, to ensure a complete solvent coverage of the complex surface making a box volume of 5,27,647 $Å^3$. To come within reach of the real situation, MD simulation is performed in water and neutralized with Na^+counter ions.

The whole system was subjected to 300 K and 1.013 bar pressure, for 20 ns at NPT ensemble of MD simulation with a recording interval of 10 ps for total energy (Kcal/mol). The stability of the protein was determined by using root mean square deviation (RMSD).

4.3.4 Acetylcholinesterase (AChE) Inhibition Assay

Begacestat, Quercetin, Ascorbic acid, Gallic acid and Caffeine were procured from Sigma Aldrich Inc, HiMedia and Loba Chemie. The molecules were tested for AChEI activity according to the method described with some modifications [30]. AChE inhibition was observed

in a 200 µl reaction mixture in a 96-well plate using Ellman's reagent or 5,5'-Dithiobis-(2-Nitrobenzoic acid (DTNB). Reaction mixture consisted of 25 µl AChE [EC 3.1.1.7] solution (25 mU in 0.1 M phosphate buffer; pH 8), 75 µl of 5,5'-dithio-bis(2-nitrobenzoic acid) (DTNB) and (0.01 M in 0.1 M phosphate buffered solution, pH 8, containing 6 mM $NaHCO_3$), 25 µl of test molecules (0.05-0.3 mM) and 50 µl of phosphate buffer (0.1 M; pH 8). The reaction mixture was incubated for 10 minutes at 37ºC. The reaction was initiated by adding 25 µl of acetylthiocholine iodide (0.075 M in water). AChE activity was determined spectrophotometrically by observing the change in absorbance for 4 min at the wavelength of 412 nm. Percent inhibition of AChE activity was calculated.

4.2.5 *In-Vitro* Neurotoxicity Assay

Neurotoxicity of the test drugs were evaluated on primary neuronal culture in 96-well plate by using 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT). For this, growth medium was replaced by serum-free DMEM having different concentration of test drugs (Quercetin, Caffeine, Ascorbic acid, Gallic acid, Donepezil, and Begacestat).

14-day old cultures were incubated with 10 µM concentration of different treatments for 6 h at 37ºC inside humidified incubator having 5% CO_2. Cells were then exposed to 25 µM $HgCl_2$ and neurons were allowed to grow further for 24 h . Cultures were washed with PBS followed by MTT assay as described above to determine cell viability and percent neuroprotection provided by drug treatments.

RESULTS AND DISCUSSION

Objective 1

Alzheimer's disease (AD) is an age-related, non-reversible brain disorder. AD ultimately leads to a severe loss of mental function. Due to lack of effective biomarkers, no effective medication available for the complete treatment of AD. There is a need to provide all AD related information to the scientific community in single platform. Our resource Alzheimer's Disease Biomarkers Comprehensive Database (ABCD) somewhat accomplished this objective by working as a big repositories of varied integrated and experimentally validated data related to AD.

Data Interface

The main ABCD interface modules are; genes search, proteins search, molecular search, gene regulatory related search and advanced Search. These allow; (i) to navigate in the database by searching genes and proteins by their name and id. In advanced search, users can search SNPs regions and mitochondrial gene in molecular search option whereas transcription factors and co-expressed gene under gene regulatory information search. Drug details, pathway catalogs and image information of AD affected brain were provided separately. Architectural diagram of ABCD shown in Figure 3.

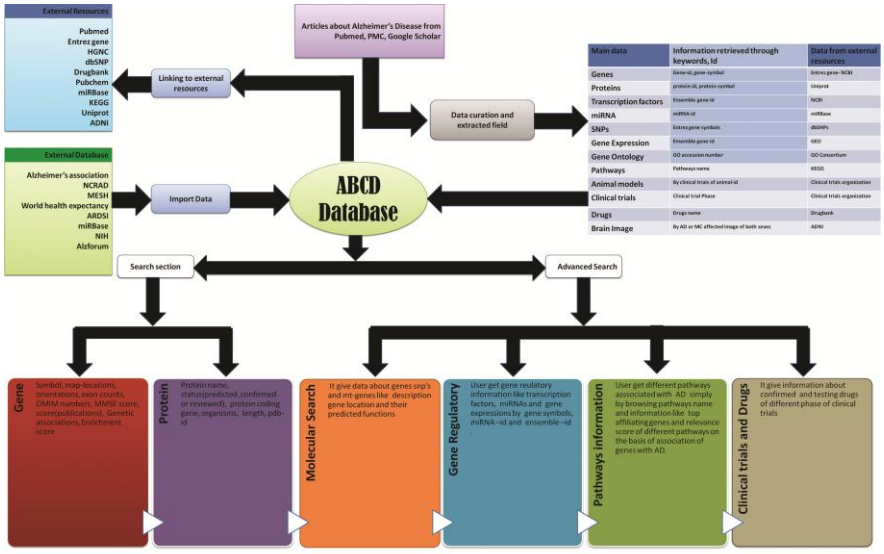

Figure 3: Architecture of ABCD

Statistics

The ABCD statistics shown in Figure 4. The number concerning the manually curated data is the following: 499 genes, 259 proteins, 66 SNPs, 404 drugs, 1608 clinical trials data etc. The data which is presented in ABCD are enriched by using external sources like alzforum, NIH, miRBASE etc. Hence, the data collected would be a unified information portal for the AD research community.

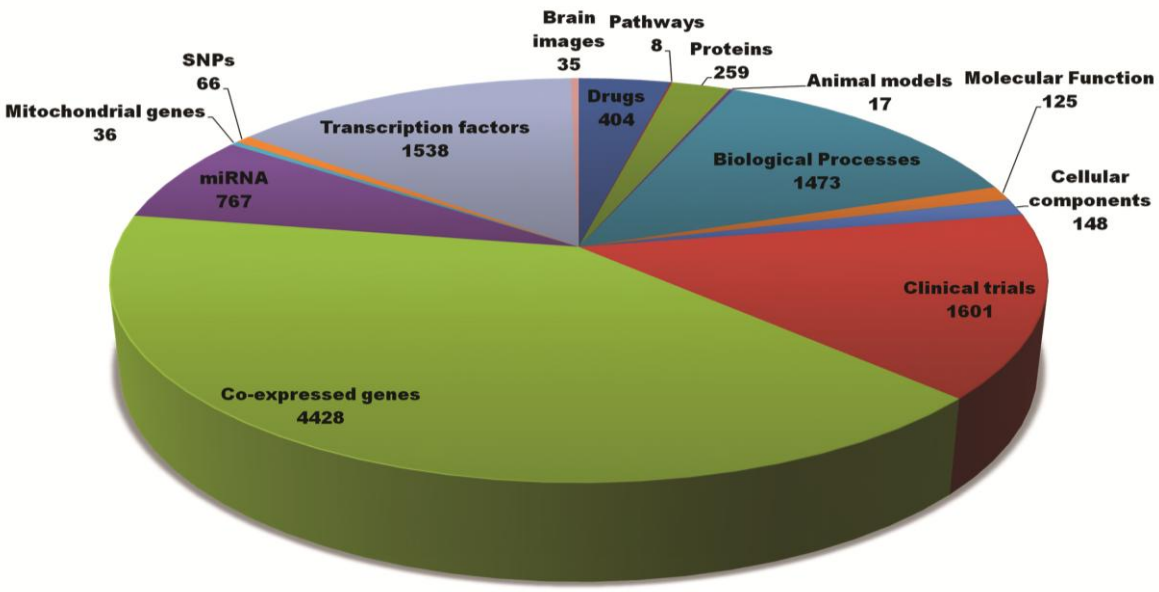

Figure 4: Statistics of different biological entities for AD stored in ABCD shown in pie chart

Objective 2

Classification of Gene set by machine learning techniques and their annotation

Analysis was performed using classification of data take place by C4.5 algorithm and J48 algorithm homologous of traditional ID3 classification algorithm. This analysis results in the correct classification of 950 out of 2111 genes. Mini mental state examination (MMSE) and GAD were the most informative variable in this geneset obtained from information gain criterion. The classification algorithm provides MMSE score cutoff value for a different stage of the disease. We generated different rules through decision tree by selecting some specific parameter at each iteration.

After preprocessing the data which involve cleaning, integration, and data reduction, the datamining was done. All analysis in this study involves 10-fold cross-validation method to test the measures without pruning the tree. A single dataset of 2111genes is used to build decision tree (DT) (Figure 5). The dataset contains information about important attributes associated with genes. In order to improve the classifier a supervised. Resample classifier was applied on data to assess the performance of the classifier, the number of tree was held constant at 100 while the

number of features was somewhat varied. Authenticity of built decision tree depends on accuracy which was measured 77% with sensitivity 98% and specificity of the classifier is 92. On applying the priori algorithm, classifier generates 10 best rules and was also verified.

Figure 5: For making the decision tree those feature are selected for classification whose values were not constant. Also we discredited the data on the basis of frequency. Different color node at the bottom of the tree reveals the class definition of leaf nodes.

Robustness of Decision Tree (DT) using C4.5 algorithm in statistical analysis tool Rapidminer

- Model of classification for different AD gene were generated on the basis of root attribute Mini-mental state examination(MMSE) whose information-gain value is high and is key attribute from where classification of geneset were taken place.
- Effectiveness of this generated tree were generated by:
  Confidence value- 0.96
  Specificity- 92%

Sensitivity- 98%

Accuracy(10-fold cross validation): 77%(Avg.)

On comparing the result, we used same dataset in another popular statistical tool for classification using J48 algorithm analogous to C4.5 Algorithm and applied test on training data of AD genes, the resulted summary result were given below:

- Number of instances: 100 (default)
- Attributes: 14
- Test:2111(Training set)
- Number of Leaves: 10
- Size of the Tree: 5
- Correctly classified instances: 93%
- Incorrectly classified instances: 7%
- Accuracy(10-fold cross validation): 77% (Avg.)

From the above observation we concluded that classification through supervised decision tree methods were popular as it is easy to understandable and perfect for handling missing values. Also it is observable that classification by two different statistical tools gives somewhat similar result and robustness of built classification model were properly maintained.

The precision or accuracy generated from this classification model for gene dataset of 2111 genes is good 77% while applying a C4.5 algorithm and J48 algorithm using different statistical tool which follow branch and bound method for classification. The absolute error rate is very low. Specificity and sensitivity are also calculated to determine the suitability of designed model and are significant. Finally, from this study it is concluded that MMSE and relevance association score are important attributes for classification of genes and labeled them to a particular class. This type of testing and analysis has been used for selection of gene for expression arrays, automated protein data annotation, automatic cancer diagnosis, plant genotype discrimination, classifying gene expression profiles and computational model for mutational sites.

Enrichment analysis of geneset at system level and topological basis

In the first step of gene ontology (GO) analysis in term of enrichment studies, the STRING interaction network tool was used as a reference to map the target seed gene-set. Out of the 761

genes, 80 were successfully mapped and used in the further analysis. XD-score helped in the identification of the closely associated pathways and cellular level processes from *KEGG* (Table1). The XD-score gives a relative average random walk distance between the entire pathway and the target genes.

Table 1: Srongly associated pathways found in *KEGG*. The XD-score determines the significance of the association. The q-value determines the significance of the overlap (Fisher test). Last column shows number of overlapping genes vs. size of the pathway.

Pathway or Process	**XD-score**	**q-value**	**Overlap/Size**
Alzheimer's disease	5.06285	0	136/138
Parkinson's disease	4.40451	0	72/99
Oxidative phosphorylation	4.38937	0	67/94
Huntington's disease	2.97584	0	82/149
Cardiac muscle contraction	1.22013	0	21/52
Amyotrophic lateral sclerosis (ALS)	0.98472	0	18/47
Long-term potentiation	0.98265	0	22/63
Long-term depression	0.50646	0.00001	13/57
GnRH signaling pathway	0.4837	0	19/83
Apoptosis	0.40646	0	19/81
Citrate cycle (TCA cycle)	0.38646	0.13586	4/26
Dorso-ventral axis formation	0.36201	0.22648	3/20
Prion diseases	0.3382	0.02727	6/35
Vascular smooth muscle contraction	0.32668	0	17/89

AD signaling pathway appeared at the top of the pathways series ranking (Figure 6), with a high XD-score of 5.06285, over three times above the significance threshold of 0.74 derived by the regression fit.

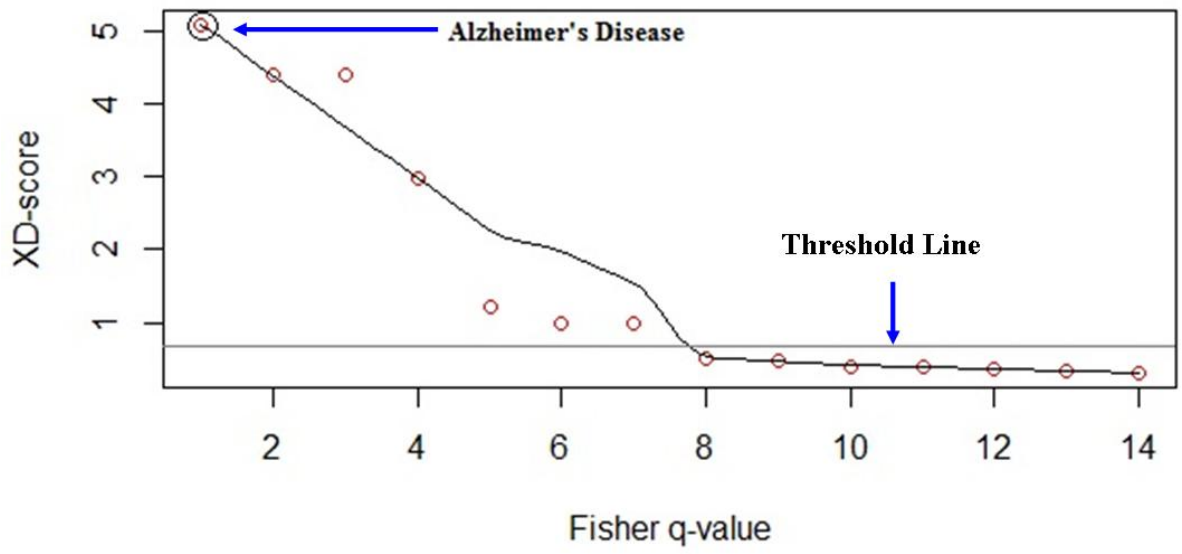

Figure 6: Regression plot between XD-score and the overlap significance (q-value) is shown. Each dot in the regression plot represents a pathway or process predicted in the enrichment analysis as related to the AD target gene set. The XD-score significance threshold is shown with a grey horizontal line. The AD pathway visible at the top had XD-score = 5.062 and q-value= 0.00.

Two important gene clusters were noticed in the network: (1) Green nodes observed in the right are specific to the related pathways (2) a mix of target set genes and pathway specific ones. The labelled orange genes in between the two clusters (*APH1B, APHIA, MMP17, and TMED10*) are the predicted path expansions. *APH1B* was already present in the input gene set and was derived again by the path expansion algorithm from the analysis of the AD pathway interactions shown in (Figure 7).

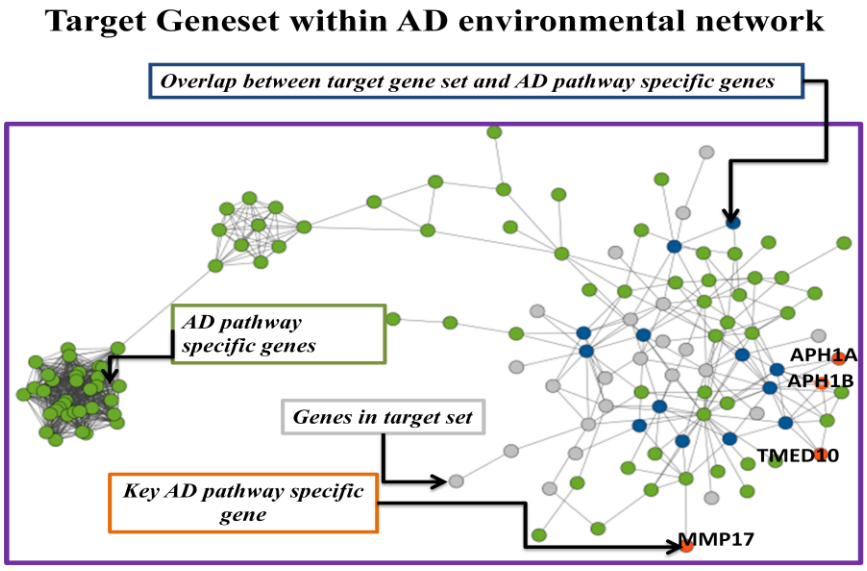

Figure 7: Target gene set within Alzheimer's disease environmental network. Node colours are: grey for the genes in the target set, blue for the overlap between the pathway and the target gene set, green for the pathway specific components and orange for pathway/process expansion. The edge colours are: green for interactions between the input set and the pathway, orange for interactions between the input set and the expansion and grey for others.

Topological Analysis

The largest connected component (80 nodes) of the enriched network was used as an input for topological analysis. Features in the topological analysis which are also present in the enriched network were then compared to those of known biological processes and pathways. Closest topological matches were obtained from KEGG database. Here we compare the topological properties by parameters of shortest path length with Node-betweenness and Degree, the graph below showed that similar pattern were found for cellular process and human diseases on network, combined influence of nodebetweenness and degree of network with shortest path length are important parametric influence which affect the behavior of associated pathways in AD network (Figure 8 A, B).

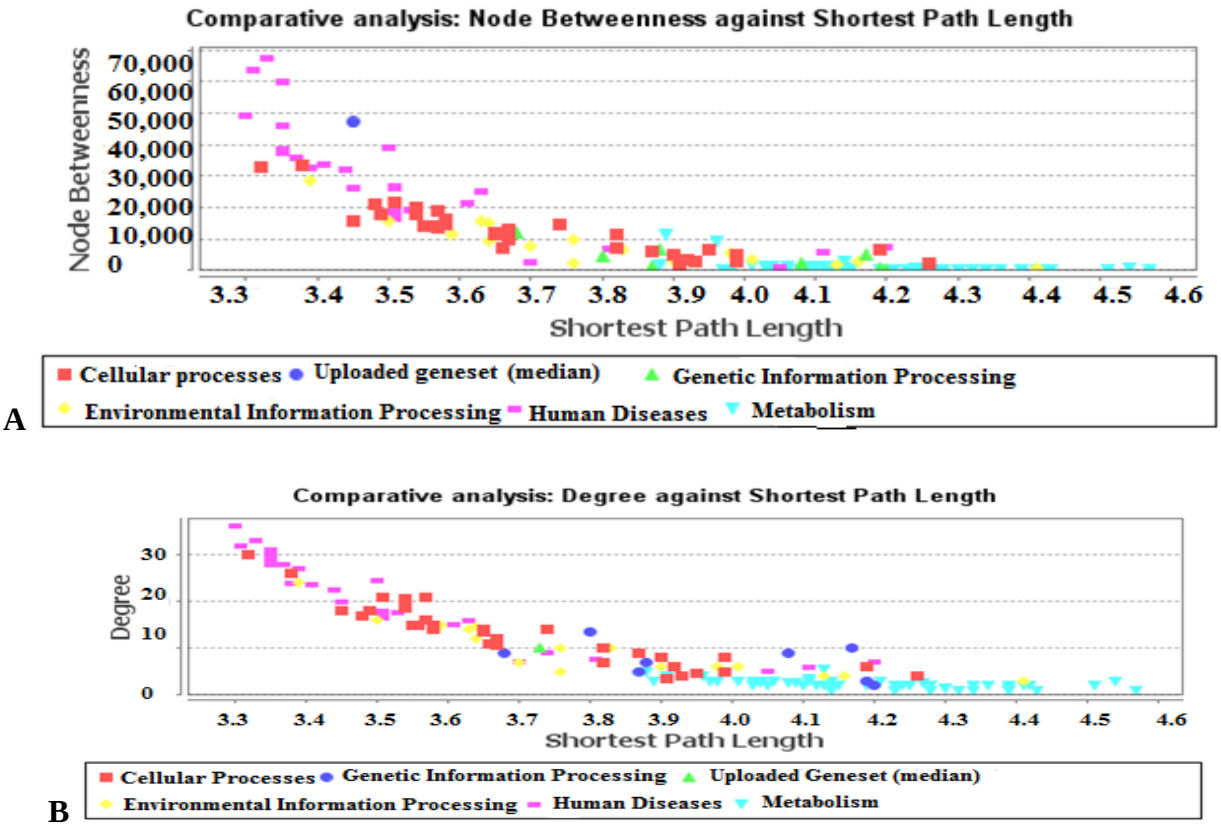

Figure 8(A): Network properties comparison between node betweenness and shortest path length; **8(B):** Network properties comparison between degree and shortest path length.

MMP17 partially modulates the differential maturation or regression of blood vessels. The topological similarity of the Parkinson's pathway (score = 0.07) is reflected by the previously found link of Parkinson's disease.

We found four prominent disease cofactors (*APH1B, APH1A, TMED10, MMP17*) using this Pathway enrichment and topological network study for candidate genes of AD. Variety of candidate genes such as NCSTN, SDHB, PICALM and their degrading capability of β-amyloid, and pathways like Parkinson's disease pathway and Oxidative phosphorylation were additionally confirmed which could be partially associative in pathological processes. Further investigations and experimental validation are require to support the hypothesis for closely connected pathways and identified disease cofactors may function a vital key to reveal the important role of those gene networks within the underlying mechanisms of the regulatory processes of AD.

Objective 3

Using AutoDock 4.2 molecular docking simulations, we predicted the various protein-ligand interactions. The different scores such as binding free energy, inhibition constant, intermolecular energy and electrostatic energy values are listed in Table 3.

Table 3: Summary of the molecular docking studies and the *in-vitro* inhibition of AChE activity

Molecules	**Binding free Energy (Kcal/mol)**	**Inhibition Constant**	**Electrostatic Energy (Kcal/mol)**	**Intermolecular Energy (Kcal/mol)**	**IC_{50} (Experimentally calculated by AChE inhibition assay)**	**Top Ranked Configuration**
Donepezil	-8.60	494.82nM	-0.14	-10.39	39.46 ± 0.75[b] nM	10
Begacestat	-6.49	17.54mM	-0.19	-8.87	104.8 ± 16.38[a] mM	4
Quercetin	-8.68	435.45nM	-0.18	-9.57	250.6 ± 4.65[b] µM	22
Ascorbic acid	-7.54	2.96mM	-1.03	-9.03	2312 ± 393.6[b] µM	16
Gallic acid	-8.46	630.93nM	-0.03	-8.76	865.8 ± 2.16[b] µM	4
Caffeine	-8.87	313.37nM	-0.29	-8.87	239.1 ± 12.95[b] µM	24

Observing the best conformation among six docked molecules, Caffeine exhibited higher number of interactions with AChE enzyme and had a binding energy of -8.87 Kcal/mol. Quercetin was ranked second in terms of binding energy of -8.68 Kcal/mol.

Molecular dynamics Simulations

To ensure the stability of the six molecules in the AChE binding site and to observe the dynamics of the docked complexes, molecular dynamics simulations of the best docked configurations of the molecules bound with AChE were performed for 20 ns using DESMOND. The time-dependent root mean square deviation (RMSD) values of the complex backbone atoms relative to the corresponding energy minimized structure were calculated. The RMSD plot and the residue fluctuation plot (RMSF) shown in Figure (8, 9) are a measure to check the stability of the protein-molecule complex.

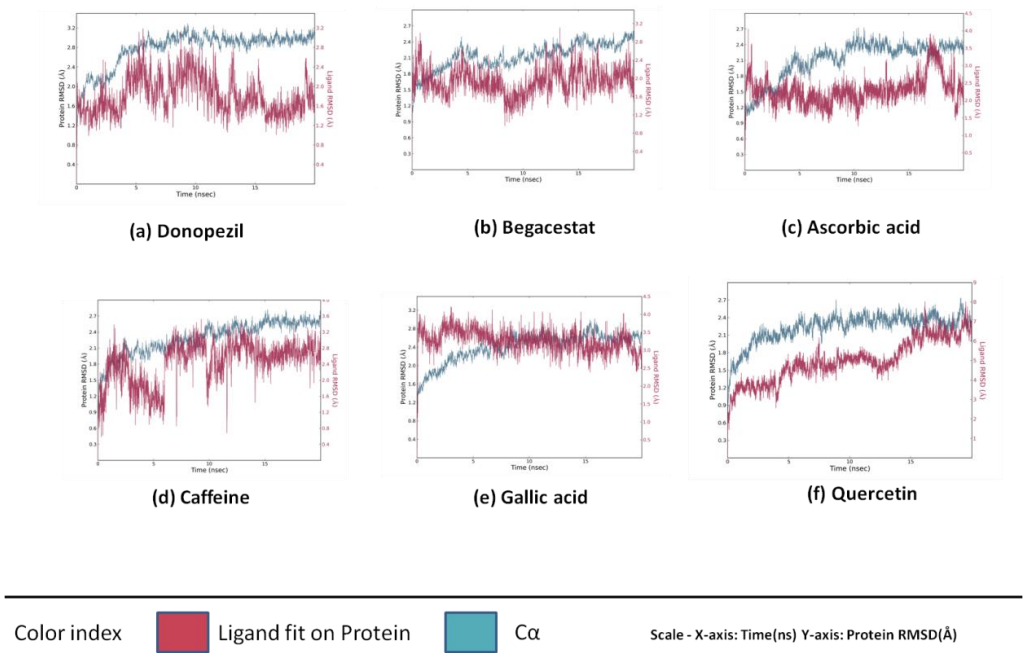

Figure 8: RMSD graph for six different small molecules

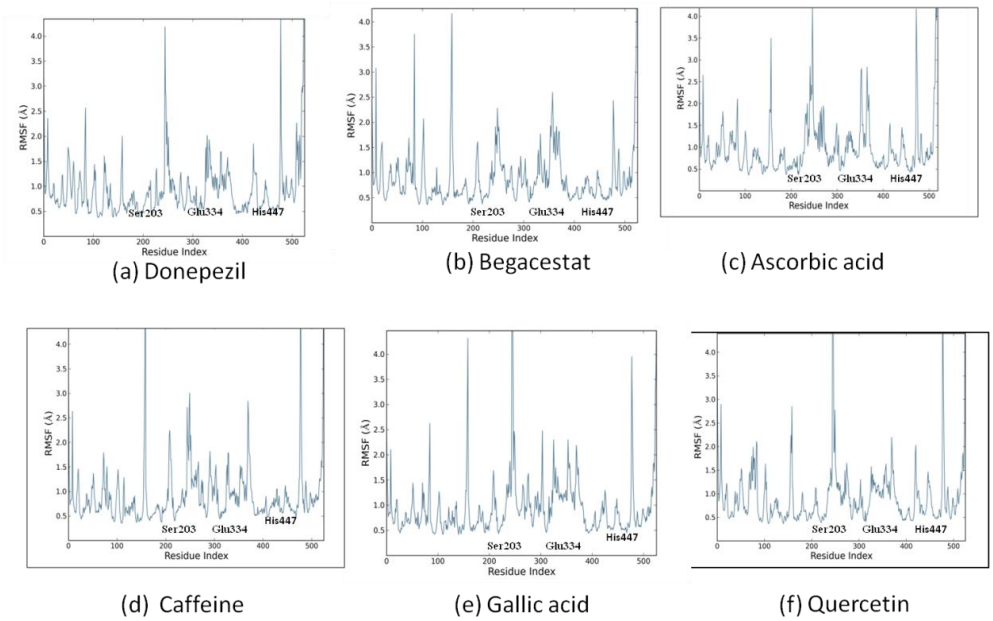

Figure 9: RMSF graph of AChE for different compounds

The RMSD of backbone atom was stabilized at 20000 ps and the ligand molecule gets stabled at 1.6 Å. The individual residues of protein fluctuate in angstrom range of 1-4 Å. Also, for AChE-Begacestat complex it was observed to achieve equilibrium around 100 ps. The structure of AChE-Caffeine complex showed deviation of acceptable value of less than 3Å RMSD with reasonably small oscillations centered on it, throughout the course of dynamics. Gallic acid and Quercetin complexes also exhibited fluctuations in the beginning of the MD simulations, however the RMSD of the ligands with protein become stable at 3Å and 7Å at 20 ns, respectively. On RMSF plot, peaks indicate areas of the protein that fluctuate the most during the simulation. Typically the tails (N and C terminal) fluctuate more than any other part of protein. Specifically by looking at the RMSF plots (Supplementary Figure S2) of the six complex structures, the active site residue Glu334 and His447 shows very small fluctuations in almost all the structures. Both the residues donate electron to Ser203 for AChE to carry the catalytic action. The total free binding energy and electrostatic interaction indicates the strong binding of molecule with the receptor.

In-vitro Acetylcholinesterase (AChE) Inhibition Assay

AChE inhibition assay is one of the most common methods used to screen molecules with potential to be implicated in AD therapeutics. Dose dependent inhibition of AChE activity was

observed for donepezil and herbal molecule and their IC_{50} values are depicted in the above given Table 3.

Donepezil showed very strong inhibition (IC_{50} = 39.46 nM) in the nanomolar range, Caffeine, Quercetin and Gallic acid showed substantial dose dependent inhibition of AChE. However, no significant inhibition of AChE activity was observed for Begacestat at the test concentration and results revealed significantly low degree of enzyme inhibition. These results suggest that quercetin, caffeine and gallic acid could be exploited for AD therapeutics, as AChEI have shown potential to delay AD onset and progression . Moreover, antioxidant and anti-inflammatory properties of these molecules are well documented [31, 32].

Neurotoxicity Assay

Herein, we used $HgCl_2$ induced neurodegeneration on primary hippocampal neuronal culture to mimic AD *in-vitro.* The neurotoxicity assay is usually carried out for drugs which may be beneficial, on neuronal cell lines or primary neuronal culture that induces Alzheimer's like neuronal damage.

$HgCl_2$ treatment inflicted high neurotoxicity (83.5 ± 8.33 %), which was significantly alleviated by treating cells with Quercetin, Caffeine, Donepezil, and Begacestat. Ascorbic acid and Gallic acid were ineffective in protecting neurons from $HgCl_2$ induced toxicity. These results indicate that Quercetin, Caffeine, Donepezil, and Begacestat possess potential rescue like activity to the neurons from neurodegeneration due to neurotoxins (Figure 10).

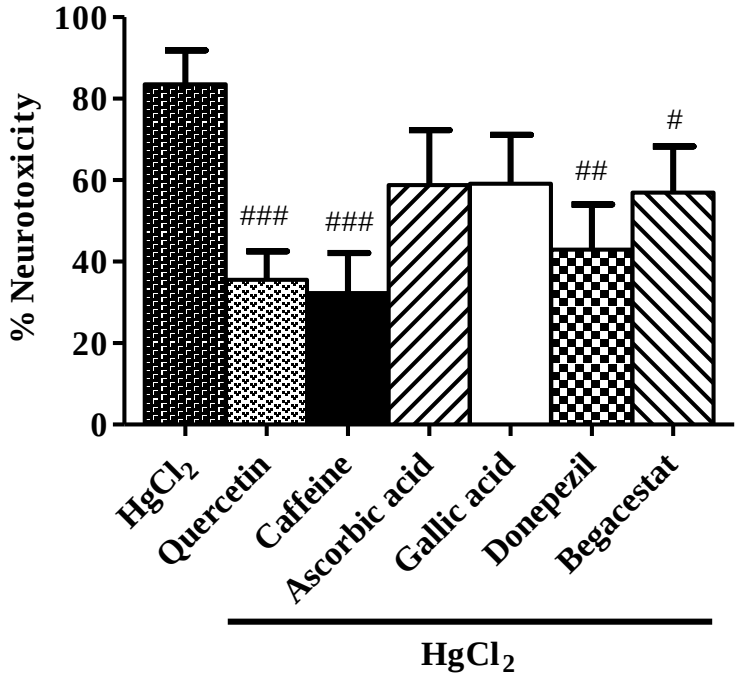

Figure 10: Effect of drug treatment on $HgCl_2$ induced neurotoxicity calculated through MTT-assay. Results are depicted as mean ± SD (n = 3).

We report that natural compounds can have disease modifying properties and modulate the cholinergic pathway, thereby, delaying the progression of AD. We analyzed the binding mechanisms of 646 small molecules to AChE using complimentary approaches including computational (molecular docking and MD simulations) and predicted quercetin, caffeine, ascorbic acid, gallic acid and begacestat to me most promising one. Results of *in-vitro* AChE inhibition assays confirmed that these molecules could be exploited as a potential therapeutic agent to inhibit AChE activity, especially quercetin and caffeine. Further, we demonstrated none of the molecules possesses neurotoxicity (10 μM) and quercetin and caffeine possesses inherited potential to protect neurons against $HgCl_2$ neurotoxin through their anti-apoptotic and anti-necrotic properties.

Conclusion

Various computational studies were performed to analyze AD at comprehensive level. The overall goal of this research was to deciphering the key biomolecules such as genes and proteins involved in AD at cellular, molecular, structural, and functional level and also find genetic association with disorder for better understanding of the complex molecular mechanism involved in AD. The overall schema and their proposed outcomes to deal with AD using various computational approaches for its system level understanding are given in Figure 11.

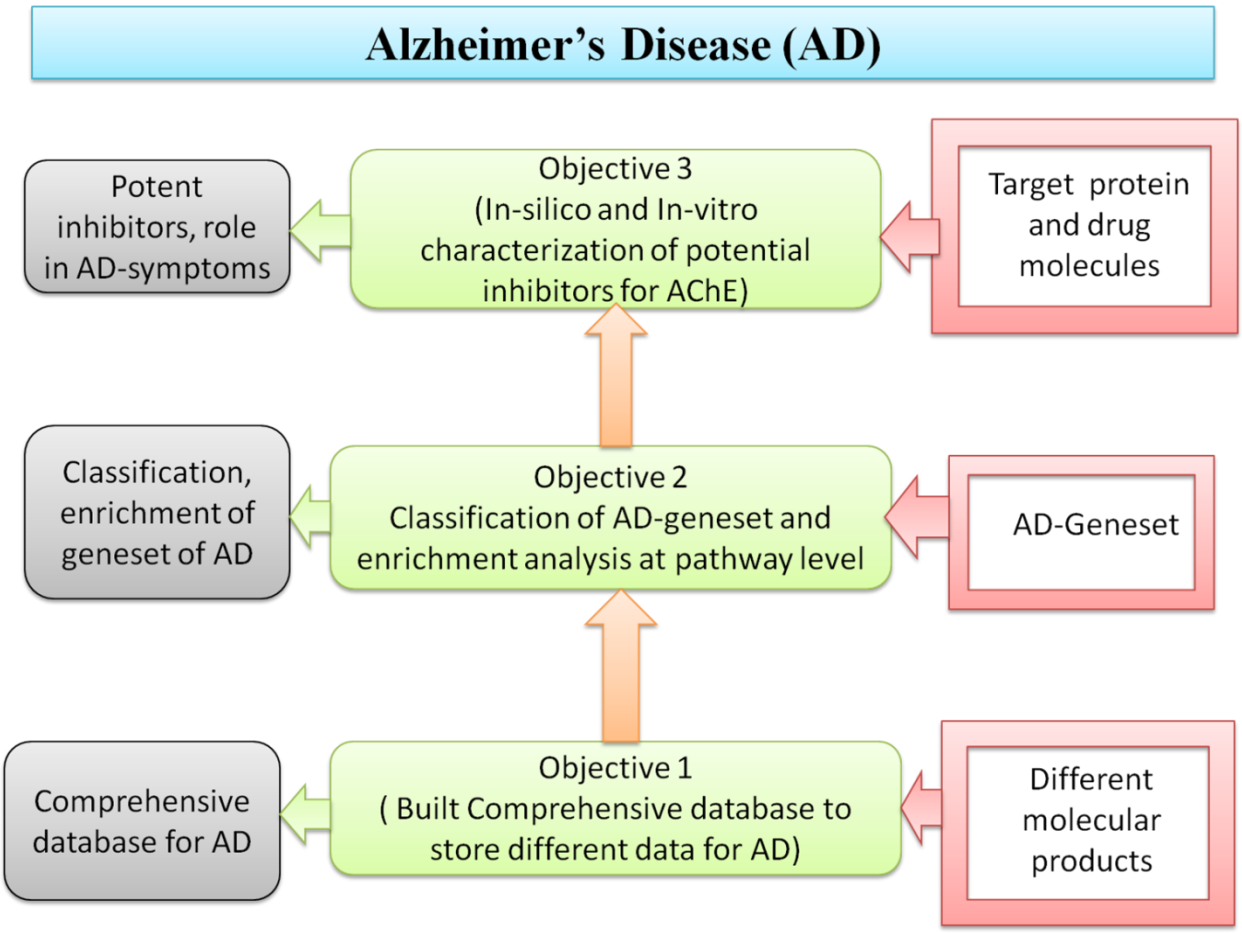
Alzheimer's Disease (AD)
Potent inhibitors, role in AD-symptoms
Objective 3 (In-silico and In-vitro characterization of potential inhibitors for AChE)
Target protein and drug molecules
Classification, enrichment of geneset of AD
Objective 2 Classification of AD-geneset and enrichment analysis at pathway level
AD-Geneset
Comprehensive database for AD
Objective 1 (Built Comprehensive database to store different data for AD)
Different molecular products

Figure 11: Overall schema of AD research objectives and their proposed outcomes specifically for assisting in early detection of disease and to provide relief for cure

It is anticipated that the proposed research will prove to be beneficial for the scientific community and the information generated will be of utmost use to the scientists and academicians for further analysis at computational and experimental levels.

References

[1] M. J. Barrett, A. F. Koeppel, J. L. Flanigan, S. D. Turner, and B. B. Worrall, "Investigation of Genetic Variants Associated with Alzheimer Disease in Parkinson Disease Cognition," *J Parkinsons Dis*, vol. 6, pp. 119-24, 2016.

[2] S. G. Mueller, N. Schuff, and M. W. Weiner, "Evaluation of treatment effects in Alzheimer's and other neurodegenerative diseases by MRI and MRS," *NMR Biomed*, vol. 19, pp. 655-68, Oct 2006.

[3] E. Cavedo, S. Lista, Z. Khachaturian, P. Aisen, P. Amouyel, K. Herholz, *et al.*, "The Road Ahead to Cure Alzheimer's Disease: Development of Biological Markers and Neuroimaging Methods for Prevention Trials Across all Stages and Target Populations," *J Prev Alzheimers Dis*, vol. 1, pp. 181-202, Dec 2014.

[4] A. Alzheimer's, "2016 Alzheimer's disease facts and figures," *Alzheimers Dement*, vol. 12, pp. 459-509, Apr 2016.

[5] P. S. Mathuranath, A. George, N. Ranjith, S. Justus, M. S. Kumar, R. Menon, *et al.*, "Incidence of Alzheimer's disease in India: a 10 years follow-up study," *Neurol India*, vol. 60, pp. 625-30, Nov-Dec 2012.

[6] A. E. Merriam, M. K. Aronson, P. Gaston, S. L. Wey, and I. Katz, "The psychiatric symptoms of Alzheimer's disease," *Journal of the American Geriatrics Society*, vol. 36, pp. 7-22, 1988.

[7] A. Serrano-Pozo, M. P. Frosch, E. Masliah, and B. T. Hyman, "Neuropathological alterations in Alzheimer disease," *Cold Spring Harb Perspect Med*, vol. 1, p. a006189, Sep 2011.

[8] K. Shimamura, N. Takatsuka, R. Inaba, H. Iwata, and H. Yoshida, "[Environmental factors possibly associated with onset of senile dementia]," *Nihon Koshu Eisei Zasshi*, vol. 45, pp. 203-12, Mar 1998.

[9] I. Piaceri, B. Nacmias, and S. Sorbi, "Genetics of familial and sporadic Alzheimer's disease," *Front Biosci (Elite Ed)*, vol. 5, pp. 167-77, Jan 01 2013.

[10] M. Kaur and P. P. Balgir, "APOE2 and consanguinity: a risky combination for Alzheimer's disease," *J Alzheimers Dis*, vol. 8, pp. 293-7, Dec 2005.

[11] D. M. Holtzman, J. Herz, and G. Bu, "Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease," *Cold Spring Harb Perspect Med*, vol. 2, p. a006312, Mar 2012.

[12] S. H. Barage and K. D. Sonawane, "Amyloid cascade hypothesis: Pathogenesis and therapeutic strategies in Alzheimer's disease," *Neuropeptides*, vol. 52, pp. 1-18, Aug 2015.

[13] S. Lista, S. E. O'Bryant, K. Blennow, B. Dubois, J. Hugon, H. Zetterberg, *et al.*, "Biomarkers in Sporadic and Familial Alzheimer's Disease," *J Alzheimers Dis*, vol. 47, pp. 291-317, 2015.

[14] J. J. Gomar, M. T. Bobes-Bascaran, C. Conejero-Goldberg, P. Davies, T. E. Goldberg, and I. Alzheimer's Disease Neuroimaging, "Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative," *Arch Gen Psychiatry*, vol. 68, pp. 961-9, Sep 2011.

[15] J. H. J. C. R. M. Beheshti, "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner," presented at the Future Generation Communication and Networking, 2008, 2008.

[16] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, pp. 2479-81, Oct 12 2004.

[17] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, *et al.*, "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic Acids Res,* vol. 35, pp. W169-75, Jul 2007.

[18] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, *et al.*, "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Res,* vol. 45, pp. D362-D368, Jan 04 2017.

[19] S. Wang, Y. Tong, T. B. Ng, L. Lao, J. K. Lam, K. Y. Zhang, *et al.*, "Network pharmacological identification of active compounds and potential actions of Erxian decoction in alleviating menopause-related symptoms," *Chin Med,* vol. 10, p. 19, 2015.

[20] K. Petrov, "Macrocyclic derivatives of 6-methyluracil: New ligands of the peripheral anionic site of acetylcholinesterase," *Int J Risk Saf Med,* vol. 27 Suppl 1, pp. S72-3, 2015.

[21] N. Furnham, G. L. Holliday, T. A. de Beer, J. O. Jacobsen, W. R. Pearson, and J. M. Thornton, "The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes," *Nucleic Acids Res,* vol. 42, pp. D485-9, Jan 2014.

[22] J. Bhachoo and T. Beuming, "Investigating Protein-Peptide Interactions Using the Schrodinger Computational Suite," *Methods Mol Biol,* vol. 1561, pp. 235-254, 2017.

[23] G. Stiglic, S. Kocbek, I. Pernek, and P. Kokol, "Comprehensive decision tree models in bioinformatics," *PLoS One,* vol. 7, p. e33812, 2012.

[24] W. A. Kukull, E. B. Larson, L. Teri, J. Bowen, W. McCormick, and M. L. Pfanschmidt, "The Mini-Mental State Examination score and the clinical diagnosis of dementia," *J Clin Epidemiol,* vol. 47, pp. 1061-7, Sep 1994.

[25] E. G. A. B. N. K. R. S. A. Valencia, "EnrichNet: network-based gene set enrichment analysis," *Bioinformatics,* vol. 28, pp. 451-457, 2012.

[26] A. C. Babtie, P. Kirk, and M. P. Stumpf, "Topological sensitivity analysis for systems biology," *Proc Natl Acad Sci U S A,* vol. 111, pp. 18507-12, Dec 30 2014.

[27] S. M. Rizvi, S. Shakil, and M. Haneef, "A simple click by click protocol to perform docking: AutoDock 4.2 made easy for non-bioinformaticians," *EXCLI J,* vol. 12, pp. 831-57, 2013.

[28] G. Kryger, M. Harel, K. Giles, L. Toker, B. Velan, A. Lazar, *et al.*, "Structures of recombinant native and E202Q mutant human acetylcholinesterase complexed with the snake-venom toxin fasciculin-II," *Acta Crystallogr D Biol Crystallogr,* vol. 56, pp. 1385-94, Nov 2000.

[29] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, *et al.*, "Integrated Modeling Program, Applied Chemical Theory (IMPACT)," *J Comput Chem,* vol. 26, pp. 1752-80, Dec 2005.

[30] A. O. Ademosun, G. Oboh, F. Bello, and P. O. Ayeni, "Antioxidative Properties and Effect of Quercetin and Its Glycosylated Form (Rutin) on Acetylcholinesterase and Butyrylcholinesterase Activities," *J Evid Based Complementary Altern Med,* vol. 21, pp. NP11-7, Oct 2016.

[31] V. Mehta, A. Parashar, A. Sharma, T. R. Singh, and M. Udayabanu, "Quercetin ameliorates chronic unpredicted stress-mediated memory dysfunction in male Swiss albino mice by attenuating insulin resistance and elevating hippocampal GLUT4 levels independent of insulin receptor expression," *Horm Behav,* vol. 89, pp. 13-22, Mar 2017.

[32] V. Mehta, A. Parashar, and M. Udayabanu, "Quercetin prevents chronic unpredictable stress induced behavioral dysfunction in mice by alleviating hippocampal oxidative and inflammatory stress," *Physiol Behav,* vol. 171, pp. 69-78, Mar 15 2017.

LIST OF PUBLICATIONS

1. **Kumar A** and Singh TR (2016) A new decision tree to solvve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system. *Interdesciplinary Science: Computational Life Science, 9(1), 1913-2751* **(I.F.: 0.8; SCI and SCOPUS indexed).**

2. **Kumar A** and Singh TR (2017) Analysis for biological network properties of Alzheimer's disease associated genesets by enrichment and topological examination. *International journal of Bioinformatics Research and Applications, 13(3), 214-22* **(SCOPUS indexed).**

3. **Kumar A**, Bansal A, Panigrahi P, Singh TR. ABCD: An Open Source Comprehensive Database for Alzheimer's disease. *Neuroinformatics* **(I.F: 2.8; Under Revision).**

4. **Kumar A**, Mehta V, Raj U, Varadwaj PK, Malairaman U, Yennamalli RM, Singh TR. Characterization of Potential Acetylcholinesterase inhibitors: Implications towards delaying the progression of Alzheimer's Disease. *Journal of Computer-Aided Molecular Design* (**I.F: 3.028; under revision).**

CONFERENCE PUBLICATIONS

1. **Ashwani Kumar**, Tiratha Raj singh: Systems biology approach for gene set enrichment and topological analysis of Alzheimer's disease pathway. International Conference on Bioinformatics and Computational Systems Biology; BSB-16, IIIT-A; 03/2016.

2. **Ashwani Kumar**, Vineet Mehta, Pritish Kumar Varadwaj, Utkarsh Raj, Udayabanu M., Tiratha Raj Singh: Flexible docking and simulation studies of herbal compounds as Acetylcholinesterase inhibitor(AChEI) for Alzheimer's disease. IAN-2015, Panjab University, Chandigarh,Punjab; 10/2015.

3. **Ashwani Kumar**, Utkarsh Raj, Pritish Kumar Varadwaj, Tiratha Raj Singh: Insights From Docking And Molecular Dynamic Simulation of Acetyl cholinesterase Inhibitors (AChEI) Structural Model For Possible Therapeutic Of Alzheimer's Disease(AD). National Conference on Bioinformatics Panorama in Agriculture and Health,2015, SHIATS, Allahabad; 10/2015.

4. **Ashwani Kumar**, Tiratha Raj Singh: A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through two standard scoring systems. National network for Mathematical and Computational Biology (NNMCB)-2014; IIT- Mandi; 10/2014.